Split leverage: attacking the confidentiality of linked databases by partitioning

Tapan Rai[1] Joanne L. Hall[2]

November 27, 2014

Abstract

This article considers the risk of disclosure in linked databases when statistical analysis of micro-data is permitted. The risk of disclosure needs to be balanced against the utility of the linked data. The current work specifically considers the disclosure risks in permitting regression analysis to be performed on linked data. A new attack based on partitioning of the database is presented.

Contents

1 Introduction **1**

2 Known attacks and existing defences **3**

3 The split leverage attack **7**

4 Extensions and Consequences of the Split Leverage Attack **9**

5 Conclusions and future directions **15**

References **16**

1 Introduction

Large amounts of micro-data are collected by government agencies through surveys, censuses and administrative sources. On the one hand, the custodians of these data are legally responsible for minimizing the risk of disclosure of sensitive information contained in these data; on the other, these data may contain vital information that may be used to inform public policy or be of other benefit to society. Useful information can be obtained by linking the micro-data collected by different data custodians. In Australia, the Australian Bureau of Statistics (ABS) is charged with the task of linking data from different custodians (e.g., government departments). In other words, the ABS serves as an integrating authority. As an integrating authority, the ABS needs to maximize the inherent value of the linked data, while protecting the legislative requirements of all data custodians. Balancing utility and disclosure risk is thus a serious issue for the ABS. Formally, disclosure is said to occur if data can be attributed to a specific entity (person or organization) from whom it was collected. On a naive level, it may appear that it would be sufficient for the ABS to adopt the same safeguards against disclosure as the submitting data custodian. However, the risk of disclosure of data from linked databases is far greater than the risk of disclosure from a single database. These risks have been studied by a number of authors including Gomatam et al [6], O'Keefe and Good [9], Reiter [10], Reiter and Kohnen [11], Reznek [12], Sparks et al (2005) [14] and Sparks et al (2008) [15]. In addition, Duncan et al [5] and Hundepool et al [7] are good general references on statistical disclosure control.

The ABS makes linked data available to legitimate users such as government departments and university researchers. To minimize the risk of disclosure, these users are not allowed direct access to the data. Instead, the users are required to log in to a secure remote server, on which they are permitted to perform specific analyses. Only the results of the analyses are made available to the user. Traditionally, the permitted analyses have been limited to the generation of descriptive statistics, graphs, and simple hypothesis tests. However due to growing demand from researchers, the ABS is working on a system that permits the development of a limited number of statistical models, such as regression analysis. Previous research on managing the disclosure risk presented by statistical modelling focused primarily on legitimate users who are not also data custodians (for example, see Sparks

(2008) et al [15] or O'Keefe and Chipperfield [8]). However, the possibility that a malicious individual working for one of the data custodians may try to obtain data to which it is not entitled adds another dimension to management of the disclosure risk. For example, one data custodian (DC-A) may develop statistical models which combine data that it contributed with data from another data custodian (DC-B). Although DC-A has complete information on the data it submitted, DC-A is not legally allowed to have access to raw data submitted by DC-B. In view of this, the ABS needs to minimize the risk that DC-A may exploit knowledge of its own data, to develop a statistical model that results in disclosure of data submitted by DC-B.

More formally, the question considered is stated as follows. Is it possible for a malicious entity to exploit statistical models to by-pass the system of safeguards that are currently in place? The specific situation considered in this work is limited to one where the malicious entity is also a data custodian. Furthermore, the statistical models considered are limited to linear regression models whose coefficients are estimated by the least squares method. Section 2 describes some of the known attacks on the system and some of the safeguards that ABS currently has in place to defend against these attacks. Section 3 presents a new attack based on partitioning the data set. Section 4 discusses some of the consequences of the new attack, and whether existing safeguards may be sufficient to protect against this attack. Section 5 provides conclusions and directions for future work.

2 Known attacks and existing defences

As a first line of defence, the ABS has established a protocol that requires users to log in to a remote server and perform statistical analysis on the server. Rather than have access to the raw data, the user is only permitted to view the results of the analysis that he or she requests. This includes descriptive statistics such as means, medians and standard deviations, graphs, results of hypothesis tests including p-values and confidence intervals, correlation and regression coefficients, their standard errors and related confidence intervals or p-values. However, even with these protections in place, the system contains vulnerabilities that can result in disclosure. Over the past decade or so, a number of researchers have studied vulnerabilities in linked statistical databases and proposed defences against them. O'Keefe

and Chipperfield [8] provides a good summary of the current state of research on reducing the risk of disclosure.

We consider a specific scenario which was presented by the ABS for consideration at MISG 2013. In this scenario the entity trying to exploit these vulnerabilities is also a data custodian who has submitted some data to the linked database. This may not be a likely scenario. However, the onus of protecting against such attacks by rogue entities or rogue employees of a genuine data custodian, falls on the ABS, in its role as an integrating authority. The scenario may be described more formally as follows.

Suppose that two data custodians, DC-A and DC-B are custodians of different sets of data for the same sample of a population. For simplicity, suppose DC-A has contributed variables $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_p$ to the linked database, and DC-B has submitted variable $\boldsymbol{y}$ to the same linked database. Since DC-A is a custodian of the variables $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_p$, it has complete information on these variables for each entity in the population. In addition, DC-A may have access to the identity of each entity in the population. However, DC-A is not legally entitled to the value of $\boldsymbol{y}$ for specific entities. Suppose also, that DC-A is entitled to perform regression analysis (via a remote server) on the linked database that includes the variables $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_p$ and $\boldsymbol{y}$. Finally, suppose that DC-A is interested in attacking the database to try and determine the value of $\boldsymbol{y}$ for a specific entity. If $\boldsymbol{y}$ is a continuous variable, some of the known vulnerabilities that DC-A could exploit are described below. We note that these vulnerabilities can be exploited by anyone with access to the remote analysis server, even if they are not data custodians. These attacks are discussed extensively in the literature (for example, see O'Keefe and Chipperfield [8]). However, the additional information available to a data custodian (DC-A) makes the system more vulnerable to these attacks. Therefore these attacks are presented below, only in the context of the attacker being the data custodian, DC-A, who has partial knowledge of the data set, as described above.

Perfect models or models with very high correlation A perfect model is one which perfectly fits the data. If DC-A is able to identify a perfect model with one or more of $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_p$ as the independent variable(s), and $\boldsymbol{y}$ as the dependent variable, it would be able to use knowledge of the variables $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_p$, to find the exact value of $\boldsymbol{y}$ for all entities in

the database. Furthermore, even if the model is not perfect, a model with very high correlation may enable DC-A to determine the value of $\boldsymbol{y}$ to a very high degree of accuracy. This would present an unacceptable level of disclosure risk. O'Keefe and Chipperfield [8] present additional details in the more general case where the attacker is not necessarily a data custodian.

Saturated Models A saturated model is one which has a very large number of independent variables. From a statistical perspective, saturated models are not considered to be good models, since they tend to underestimate the regression coefficients of individual independent variables. However, saturated models often have very high predictive ability, and can generally be used to predict the value of the dependent variable with a high degree of accuracy. Once again, if DC-A is able to exploit this property of saturated models, then it would present an unacceptable disclosure risk. O'Keefe and Chipperfield [8] and Ritchie [13] present additional details in the general case where the attacker is not necessarily a data custodian.

Sparse Models In this context, a sparse model is one which has very few unique data points. For example a sparse model may consist of just one data point, in which case, it presents the exact value for a specific entity. Alternatively, if the number of data points equals the number of independent variables, the model equation could exactly (or with a very high degree of accuracy) determine the hyperplane through the points that were used to develop the model. From a statistical perspective, sparse models are not considered good models, because they rarely (if ever) have statistically significant coefficients. However, if DC-A is able to fit a sparse model, it would be able to accurately identify the value of $\boldsymbol{y}$ for each entity, whose data was used to develop the regression model. O'Keefe and Chipperfield [8] and Sparks et al (2008) [15] present details of the issues related sparse models in a more general context.

There are some obvious defences against the simple vulnerabilities described above. These have been studied extensively, and have been implemented in the ABS remote server systems. Chipperfield et al [2]present these and other simple defences. Some of these defences include the following.

Models with very high correlation are not permitted The ABS system is designed to disallow regression models with $R^2 > 0.95$ $(R > 0.975)$. This provides a defence against the vulnerability presented by perfect models or models with very high correlation.

Defence against saturated models The number of independent variables is restricted to 30. This reduces the likelihood that any user is able to develop a saturated model.

Defence against Sparse models The minimum number of points that can be selected for modeling is set at 50. Together with the restriction on the number of independent variables, this prevents a user from developing a sparse model.

Several other simple attacks have been considered and defences against these attacks have already been implemented in the ABS system. These attacks and defences are not considered here. Instead, the rest of this section focuses on some of the more sophisticated known attacks, which involve comparing or aggregating the results of several regression models obtained from the remote server.

Differencing Attack Suppose DC-A wants to target the y-value, y_E, associated with a specific entity, E. DC-A may try to achieve this as follows: First, DC-A uses the remote server to develop a regression model on the entire data set with $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_p$ as the independent variables and $\boldsymbol{y}$ as the dependent variable. Suppose this regression yields $\beta_0, \beta_1, \beta_2, \ldots, \beta_p$ as the estimates of the regression coefficients. Next, DC-A drops the data point $(x_{1E}, x_{2E}, \ldots, x_{pE}, y_E)$ related to E, and performs the same regression on the reduced data set. Suppose the corresponding estimates of the regression coefficients are $\beta_{0E'}, \beta_{1E'}, \ldots, \beta_{pE'}$. DC-A can then exploit its knowledge of $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_p$, to calculate y_E. Cox [3] and in O'Keefe and Chipperfield [8] present the differencing attack in a more general context.

Leverage Attack: The leverage h of a data point $(x_1, x_2, \ldots, x_p)$ is a measure of its distance from $(\bar{\boldsymbol{x}}_1, \bar{\boldsymbol{x}}_2, \ldots, \bar{\boldsymbol{x}}_p)$ where $\bar{\boldsymbol{x}}_i$ is the arithmetic mean of the variable $\boldsymbol{x}_i$ [16]. If a data point has a high leverage value, a regression model may predict its y-value very accurately. The leverage of a point

depends only on its values on the independent variables and not on the dependent variable. Therefore, DC-A can use its knowledge of $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_p$ to identify leverage points and exploit this information to determine the y-value for any entity that has a high leverage value. On a more sophisticated level, DC-A could use an appropriate transformation on $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_p$ to force a specific entity to have a high leverage value. It could then use the remote server to develop a regression model with the transformed $\boldsymbol{x}$-variables as the independent variables and $\boldsymbol{y}$ as the dependent variable. The resulting regression equation can be expected to accurately determine the y-value for the entity with high leverage value. Details of the leverage attack are presented in a more general context by O'Keele and Chipperfield [8] and Gomatam et al [6].

The ABS has implemented a number of defences against these and other potentially sophisticated attacks on the system. The simpler defences involve dropping points from the data set and restricting transformations. For example, points with a high leverage value are dropped from the requested analysis, if they are detected. In addition, a few points are randomly dropped from the data set to minimize the risk of a successful attack against a specific entity. Furthermore there are restrictions on the types of transformations that are permitted. Specifically, transformations that combine variables that are contributed by different data custodians are disallowed. All these defences are designed to prevent leverage and differencing attacks, and to minimize the likelihood of success of other attacks that may not yet be known. Chipperfield et al [2] present details of these defences.

A more sophisticated defence involves perturbation of the regression output. This defence mechanism works by adding a small amount of noise to the estimating equations. The amount of noise needs to be carefully determined since large changes in the size of the regression coefficients or related p-values would compromise the utility of the output. In addition, the perturbation cannot significantly alter the error distribution, since it could compromise confidence in the model fit. A detailed anslysis of these issues is presented by Dwork et al [4]. The perturbation algorithm used by the ABS is confidential and is not presented here. Some of the details and associated challenges are presented by Chipperfield at al [2].

3 The split leverage attack

This section presents a new attack called the *split leverage attack*, which was developed during the MISG. The attack is designed to exploit the vulnerability presented by the existence of a high leverage point in the dataset. It involves partitioning the data into disjoint subsets to prevent the system from detecting the leverage point; hence the name, *split leverage attack*. In general, finding an appropriate partition may be a difficult task, especially if the attacker does not have access to the raw data. However, if the attacker is also the data custodian who submitted the raw data on the independent variables, then finding an appropriate partition would certainly be feasible. This is a key assumption of the attack.

As in the previous section, suppose that two data custodians, DC-A and DC-B are custodians of different sets of data for the same sample of a population. For simplicity, suppose DC-A has contributed data on variables $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_p$ to the linked database, and DC-B has submitted data on a continuous variable $\boldsymbol{y}$ to the same linked database. Since DC-A is a custodian of the variables $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_p$, it has complete information on these variables for each entity in the population. However, DC-A is not legally entitled to the value of $\boldsymbol{y}$ for specific entities. Suppose also, that the data set contains a high leverage point, l and that DC-A is interested in exploiting knowledge of this high leverage point to gain information about the $\boldsymbol{y}$-value of the entity associated with l. As before, assume that DC-A is permitted to perform regression analysis (via a remote server) on the linked database that includes the variables $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_p$ and $\boldsymbol{y}$. DC-A cannot directly use the remote server to gain information of the y-value associated with l, since the server would detect l as a high leverage point and exclude it from the analysis. The split leverage attack, which is designed to bypass this protection, is described below.

Proposition 1 (Split leverage attack) *1. Assume that* DC-A *has complete information on the data matrix,*

$$X = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix},$$

associated with the independent variables, $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_p$.

2. *Assume that* X *contains a high leverage point* l*, and that* DC-A *uses its knowledge of complete information on* X *to find* l.

3. *Next, suppose* DC-A *partitions the data matrix,* X*, into two disjoint subsets* X_1 *consisting of* m *rows of* X *and* X_2*, consisting of the remaining* $(n-m)$ *rows of* X*. Both* X_1 *and* X_2 *contain data on all* p *independent variables; however,* X_1 *and* X_2 *contain different data points.*

4. *Suppose also, that* DC-A *is able exploit complete information on* X *to find* X_1, X_2 *such that* $X = X_1 \cup X_2$*,* $X_1 \cap X_2 = \emptyset$*, and neither* X_1 *nor* X_2 *contains a high leverage point.* DC-A *can also ensure that each of these subsets contains the minimum number of points required to circumvent other protections.*

5. DC-A *can then use the remote server to separately regress* $\boldsymbol{y}$ *on* X_1 *and* $\boldsymbol{y}$ *on* X_2.

6. *Let* $\beta_{01}, \beta_{11}, \beta_{21}, \ldots, \beta_{p1}$ *be the estimates of the regression coefficients obtained by regressing* $\boldsymbol{y}$ *on* X_1 *and* $\beta_{02}, \beta_{12}, \beta_{22}, \ldots, \beta_{p2}$ *be the estimates of the regression coefficients obtained by regressing* $\boldsymbol{y}$ *on* X_2.

7. *Next, suppose* DC-A *uses its own computer to calculate* $\hat{y}_{i1} = \beta_{01} + \beta_{11}x_{i1} + \beta_{21}x_{i2} + \cdots + \beta_{p1}x_{ip}$ *for each* $(x_{i1}, x_{i2}, \ldots, x_{ip}) \in X_1$ *and* $\hat{y}_{j2} = \beta_{02} + \beta_{12}x_{j1} + \beta_{22}x_{j2} + \cdots + \beta_{p2}x_{jp}$ *for each* $(x_{j1}, x_{j2}, \ldots, x_{jp}) \in X_2$.

8. *Let* $\hat{\boldsymbol{y}}_1 = \{\hat{y}_{i1} : x_i \in X_1\}$*,* $\hat{\boldsymbol{y}}_2 = \{\hat{y}_{j2} : x_j \in X_2\}$ *and* $\hat{\boldsymbol{y}} = \hat{\boldsymbol{y}}_1 \cup \hat{\boldsymbol{y}}_2$.

9. DC-A *can then use complete information on* X*, and its own computer to regress the* $\hat{\boldsymbol{y}}$ *values on the entire set* X.

10. *Then, the values of the regression coefficients of this final regression model of* $\hat{\boldsymbol{y}}$ *on* X *are identical to those obtained by regressing* $\boldsymbol{y}$ *on* X.

11. DC-A *can thus overcome the protection against high leverage points and obtain a regression model with a high leverage point* l.

12. *Since* l *is a high leverage point, the model accurately estimates the value of* $\boldsymbol{y}$ *associated with* l*, resulting in unacceptable disclosure.*

The proposition (specifically point 10) follows from a more general result (Theorem 2), which is proved in Section 4. The viability and consequences of this attack are also discussed in the next section.

4 Extensions and Consequences of the Split Leverage Attack

The ABS system has a number of protections in place to minimize the risk of disclosure in various situations. The obvious question, then, is whether any of the existing protections might be sufficient to protect against the split leverage attack presented in the previous section. The most interesting question in this regard is whether perturbation of the regression coefficients might have a protective effect against the split leverage attack. This question arose at MISG, but is yet to be considered. Another interesting question is whether the measure of disallowing models with high correlation coefficient ($R^2 > 0.95$) might be sufficient to protect against the split leverage attack. This question was also not considered at the MISG. However, subsequent simulations were performed using randomly generated datasets. The regression coefficients for the randomly generated data and the split leverage attack were estimated using Mathematica 8 [17]. An examination of the results of these simulations suggested the following.

1. The regression on the original data set X, can be re-created (to a very high degree of precision) by the split leverage attack provided that the subsets X_1 and X_2, into which X is partitioned, are disjoint.

2. If X contains a high leverage point, then it is possible to partition the original data set into disjoint subsets X_1 and X_2, neither of which contains a high leverage point, provided $|X| \geq 6$.

3. The split leverage attack appears to work even for data where the correlation coefficient (on the full data set) is below 0.8.

4. The attack also appears to work when the correlation coefficient on each of the subsets X_1 and X_2 is low (of the order of 0.3 or below).

5. Some simulations showed that the correlation coefficients of the regression models on the disjoint subsets X_1 and X_2 can be much lower than on the original regression model on X.

6. The re-creation of the regression on the full data set appears to work whether the full data set has a leverage point or not.

Point 3, above, suggests that the protection against models with high correlation is not sufficient to protect against the split leverage attack. In addition, points 3, 4, 5 and 6 suggest that it might be possible to overcome the protection against models with high correlation coefficient, by partitioning the data into disjoint subsets. However, partitioning the data into subsets that lowers the correlation coefficient for models on both subsets X_1 and X_2 may be hard to achieve since it may require some information about the y-values. Of course, the attacker may try to use brute force to find an appropriate partition. It is unlikely that a brute force attack of this nature would be computationally feasible. On the other hand, the attacker may be able to gain partial information about the y-values through some kind of intelligent sequential partitioning of the dataset. Such an attack has not yet been explored. Point 6 is formally proved in the following theorem. Point 1 (the split leverage attack) follows as a special case, in which the data set contains a leverage point. Although points 3, 4 and 5 are not explicitly proved, the follow implicitly from the theorem. The formal exploration of point 2 remains an open conjecture.

Theorem 2 *Let X, Y be continuous random variables. The linear regression of Y on X can be re-created as follows:*

1. *Partition X into two disjoint sets X_1 and X_2. Let Y_1 be the y-values associated with the x-values in X_1 and Y_2 be the y-values associated with x-values in X_2.*

2. *Perform regression on (X_1, Y_1); let $\hat{Y}_1$ be the y-values predicted by this regression.*

3. *Perform regression on (X_2, Y_2); let $\hat{Y}_2$ be the y-values predicted by this regression.*

4. *Let $(X, \hat{Y}) = (X_1, \hat{Y}_1) \cup (X_2, \hat{Y}_2)$.*

5. Perform a linear regression on $(X, \hat{Y})$.

Then, regressing $\hat{Y}$ *on* X *yields the same equation as regressing* Y *on* X.

This result is a little more general than the split leverage attack, since it does not require the presence of a high leverage point. This could be an important result in information security, since it shows how to reconstruct a regression model on a data set, by developing models on subsets of the data. However, it may not be of much interest to statisticians since it does not appear to have any direct application to statistics. In spite of this, Wetherill [16] presents a number of results similar to this one, in a section on sub-model analysis. If these results on sub-model analysis could be re-interpreted in the context of information security, then they could potentially be used to refine and improve the split leverage attack, or to develop new attacks on linked databases.

We present the proof of Theorem 2 below, and refer the reader to Anton and Rorres [1] or any other standard book for the details of the linear algebra.

Proof: Let $X = \{\boldsymbol{x}_1^T, \boldsymbol{x}_2^T, \ldots, \boldsymbol{x}_n^T\}$ be a set of row vectors of length m, and let Y be represented by $\boldsymbol{y}$, a vector of length n. Let M be the matrix

$$M = \begin{bmatrix} 1 & \boldsymbol{x}_1^T \\ 1 & \boldsymbol{x}_2^T \\ \vdots & \vdots \\ 1 & \boldsymbol{x}_n^T \end{bmatrix}.$$

Then the coefficients of the least squares regression of Y on X are

$$\boldsymbol{v} = (M^T M)^{-1} M^T \boldsymbol{y}. \tag{1}$$

Suppose that the set $N = \{1, 2, \ldots, n\}$ is partitioned such that $A \cup B = N$, $A \cap B = \emptyset$ where $X_1 = \{\boldsymbol{x}_i \mid i \in A\}$ and $X_2 = \{\boldsymbol{x}_i \mid I \in B\}$. Let

$$P = \begin{bmatrix} 1 & \boldsymbol{x}_i^T \\ 1 & \boldsymbol{x}_j^T \\ \vdots & \vdots \\ 1 & \boldsymbol{x}_k^T \end{bmatrix} \quad \text{such that } i < j < k \in A,$$

$$Q = \begin{bmatrix} 1 & \boldsymbol{x}_i^T \\ 1 & \boldsymbol{x}_j^T \\ \vdots & \vdots \\ 1 & \boldsymbol{x}_k^T \end{bmatrix} \quad \text{such that } i < j < k \in B.$$

Let Y_1 be represented by the vector

$$\boldsymbol{\phi} = \begin{bmatrix} y_i \\ y_j \\ \vdots \\ y_k \end{bmatrix} \quad \text{such that } i < j < k \in A,$$

and Y_2 be represented by the vector

$$\boldsymbol{\psi} = \begin{bmatrix} y_i \\ y_j \\ \vdots \\ y_k \end{bmatrix} \quad \text{such that } i < j < k \in B.$$

Then the coefficients of the least squares regression of Y_1 on X_1 are

$$\boldsymbol{v}_A = (P^T P)^{-1} P^T \boldsymbol{\phi},$$

and the coefficients of the least squares regression of Y_2 on X_2 are

$$\boldsymbol{v}_B = (Q^T Q)^{-1} Q^T \boldsymbol{\psi}. \tag{2}$$

Let $\hat{Y}$ be represented by the vector $\boldsymbol{\zeta}$ such that

$$\zeta_i = \begin{cases} M_i \cdot \boldsymbol{v}_A & \text{if } i \in A\,, \\ M_i \cdot \boldsymbol{v}_B & \text{if } i \in B\,, \end{cases}$$

where M_i is the ith row of M.

Without loss of generality assume that the rows of M and corresponding entries of $\boldsymbol{y}$ have been arranged such that

$$M = \begin{bmatrix} P \\ Q \end{bmatrix} \quad \text{and} \quad \boldsymbol{y} = \begin{bmatrix} \boldsymbol{\phi} \\ \boldsymbol{\psi} \end{bmatrix}.$$

Consequently

$$\zeta = \begin{bmatrix} P\boldsymbol{v}_A \\ Q\boldsymbol{v}_B \end{bmatrix}$$

Note that

$$M^T\boldsymbol{\zeta} = \begin{bmatrix} P^T Q^T \end{bmatrix} \begin{bmatrix} P\boldsymbol{v}_A \\ Q\boldsymbol{v}_B \end{bmatrix} = \begin{bmatrix} P^T P\boldsymbol{v}_a + Q^T Q\boldsymbol{v}_B \end{bmatrix} \tag{3}$$

and

$$M^T\boldsymbol{y} = \begin{bmatrix} P^T Q^T \end{bmatrix} \begin{bmatrix} \boldsymbol{\psi} \\ \boldsymbol{\phi} \end{bmatrix} = \begin{bmatrix} P^T\boldsymbol{\phi} + Q^T\boldsymbol{\psi} \end{bmatrix}. \tag{4}$$

Consider $\boldsymbol{\phi}$, then

$$\boldsymbol{v}_A = (P^T P)^{-1} P^T \boldsymbol{\phi}, \tag{5}$$

$$(P^T P)^{-1} P^T P\boldsymbol{v}_A = (P^T P)^{-1} P^T \boldsymbol{\phi}, \tag{6}$$

$$P^T P\boldsymbol{v}_A = P^T \boldsymbol{\phi}. \tag{7}$$

A similar argument shows that

$$Q^T Q\boldsymbol{v}_B = Q^T \boldsymbol{\psi}. \tag{8}$$

From equations (3), (7) and (8)

$$M^T\boldsymbol{\zeta} = \begin{bmatrix} P^T\boldsymbol{\phi} + Q^T\boldsymbol{\psi} \end{bmatrix} = M^T\boldsymbol{y}. \tag{9}$$

Hence

$$(M^T M)^{-1} M^T\boldsymbol{\zeta} = (M^T M)^{-1} M^T\boldsymbol{y} = \boldsymbol{v}. \tag{10}$$

Hence the least squares regression equation of $\hat{Y}$ on X is identical to the least squares regression equation of Y on X. ♠

We conclude this section with a discussion of possible defences against the split leverage attack. The split leverage attack will only re-create regression equations—not the original data. So if the regression output from the server is perturbed, then it would only re-create the perturbed y-values. In theory, this suggests that perturbation may provide some protection against this

attack. However, even when the regression coefficients are perturbed, the regression model can be expected to predict the y-value for a high leverage point with a high level of accuracy. That is why regressions with leverage points are disallowed.

Since it is believed that protection against leverage attacks requires a measure in addition to perturbation, it is unlikely that perturbation alone will be sufficient to protect against the split leverage attack, which bypasses the existing protection against leverage attacks. On the other hand, it is possible that the dropping of random data points from all regressions (as a protection measure) could provide some protection against the split leverage attack, since it could result in the high leverage point being dropped from the subsets X_1 and X_2. Alternatively, since the attack requires disjoint sets, it may be possible to prevent the attack by randomly adding a few data points to all regressions performed on the server. The data points would need to be carefully selected to have x-values that are close to the ones in the data set that the attacker/researcher is interested in. Another alternative worth exploring might involve adding a few phantom data points (not real data) to make marginal changes to the regression coefficients in all analyses. The addition of these phantom data points could have the effect of preventing the partitioning of the data into disjoint sets. Of course, if an attacker is able to find an appropriate way to adjust for the fact that $X_1 \cap X_2 \neq \emptyset$, then the addition of data points may not be very useful.

5 Conclusions and future directions

As with all information security problems, protecting against disclosure is likely to be a cat and mouse game between the integrating authority and the attacker. Even if the integrating authority has unlimited resources, it may be impossible to anticipate all possible methods of attack and to provide protections against them. The MISG was able to provide the ABS with some new issues to consider in improving the security of their system. These include:

- a new attack that exposes a potential vulnerability in the system;
- considering various methods to protect against the new attack;

Consideration of these issues present several new directions for future work:

- a formal investigation of the structure of data sets which contain a leverage point, that can be partitioned into subsets which do not contain leverage points;
- examining whether a dataset can be systematically partitioned to overcome the defence against models with high correlation;
- determining whether the existing defences are sufficient to protecting against the split leverage attack;
- developing new protections against the split leverage attack.

Acknowledgements We are grateful to ABS representatives Daniel Elazar and James Chipperfield for introducing us to this problem. We also acknowledge the contribution of the other members of the team: Tony Pettitt, Douglas Stebila, Scott Alexander, Jacobien Carstens and Anagi Gamachchi. In particular, we thank Daniel Elazar for leading the discussion of the problem at MISG 2013 and Tony Pettitt for providing comments on the draft of this article.

References

[1] Anton, H., Rorres, C. *Elementary Linear Algebra*, 10th edition. John Wiley and Sons Inc, Hoboken, NJ, 2010.

[2] Chipperfield, J. O., Yu, F., Gare, M. Providing access to microdata for statistical purposes: Experiences of the Australian Bureau of Statistics with remote analysis servers. *Symposium 2011Catalogue no. 11-522-XCB, Statistics Canada*, pp 187-194, http://publications.gc.ca/collections/collection_2013/statcan/11-522-x/CS11-522-2011-eng.pdf, 2011.

[3] Cox, L. Confidentiality Issues For Statistical Database Query Systems. *Invited Paper for Joint UNECE/Eurostat Seminar on Integrated*

Statistical Information Systems and Related Matters, Geneva Switzerland, http://www.unece.org/stats/documents/ces/sem, 2002.

[4] Dwork, C., McSherry, F., Nissim, K., Smith, A. Calibrating noise to sensitivity in private data analysis. *Proceedings of the 3rd Theory of Cryptography Conference*, LNCS 3876, pp 265284, 2006. doi: 10.1007/11681878_14

[5] Duncan, G. T., Elliott, M., Salazar-Gonzàlez, J.-J. *Statistical Confidentiality: Principles and Practice*. Springer, NY, 2012. doi: 10.1007/978-1-4419-7802-8

[6] Gomatam, S., Karr, A., Reiter, J., Sanil, A. Data dissemination and disclosure limitation in a world without microdata: A risk-utility framework for remote access systems. *Statistical Science* 20(2), pp 163-177, 2005. doi: 10.1214/088342305000000043

[7] Hundepool, A., Domingo-Ferrer, J., Franconi, L., Giessing, S., Norholdt, E., Spicer, K., de Wolf, P.-P. *Statistical Disclosure Control*. Wiley, UK, 2012.

[8] O'Keefe, C., Chipperfield, J. A Summary of Attack Methods and Confidentiality Protection Measures for Fully Automated Remote Analysis Systems. *International Statistical Review* 81(3), pp 426-455, 2013. doi: 10.1111/insr.12021

[9] O'Keefe, C., Good, N. Regression output from a remote analysis server. *Data and Knowledge Engineering* 68(11), pp 1175-1186, 2009. doi: 10.1016/j.datak.2009.06.009

[10] Reiter, J. P. Model diagnostics for remote-access regression servers. *Statistics and Computing* 13(4), pp 371380, 2003. doi: 10.1023/A:1025623108012

[11] Reiter, J. P., Kohnen, C. N. Categorical data regression diagnostics for remote servers. In *Journal of Statistical Computation and Simulation* 75(11), pp 889903, 2005. doi: 10.1080/00949650412331299184

[12] Reznek, A. P. Recent confidentiality research related to access to enterprise microdata. *Comparative Analysis of Enterprise Microdata*

(CAED) Conference Chicago, IL, USA, http://www.oecd.org/std/37503027.pdf, 2006.

[13] Ritchie, F. Disclosure Controls for Regression Outputs. *L*ondon: Mimeo, Office of National Statistics, London, http://www.wiserd.ac.uk/files/7913/6543/6668/WISERD_WDR_006.pdf, 2006.

[14] Sparks, R., Carter, C., Donnelly, J., Duncan, J., O'Keefe, C., Ryan, L. A framework for performing statistical analyses of unit record health data without violating either privacy or confidentiality of individuals. *Proceedings of the 55th Session of the International Statistical Institute,* Sydney, 2005.

[15] Sparks, R., Carter, C., Donnelly, J. B., O'Keefe, C., Duncan, J., Keighley, T., McAullay, D. Remote access methods for exploratory data analysis and statistical modelling: Privacy-Preserving Analytics. *Computer Methods and Programs in Biomedicine* 91(3), pp 208-222, 2008. doi: 10.1016/j.cmpb.2008.04.001

[16] Wetherill, G. B. *Regression Analysis with Applications*. Chapman and Hall Ltd, London, 1986.

[17] Wolfram Research, Inc. *Mathematica* Version 8.0. Wolfram Research, Inc., Champaign, IL, USA, 2010.

Author addresses

1. **Tapan Rai**, University of Technology Sydney, School of Mathematical Sciences, Sydney, Australia.
`mailto:Tapan.Rai@uts.edu.au`

2. **Joanne L. Hall**, Queensland University of Technology, Mathematical Sciences School, Brisbane, Australia.
`mailto:j42.hall@qut.edu.au`